Estimation of alternative splicing isoform frequencies from RNA-Seq data

Marius Nicolae[1], Serghei Mangul[2], Ion Măndoiu[1], and Alex Zelikovsky[2]

[1] Computer Science & Engineering Department, University of Connecticut
371 Fairfield Way, Storrs, CT 06269
{man09004,ion}@engr.uconn.edu
[2] Computer Science Department, Georgia State University
University Plaza, Atlanta, Georgia 30303
{serghei,alexz}@cs.gsu.edu

Ubiquitous regulatory mechanisms such as the use of alternative transcription start and polyadenylation sites, alternative splicing, and RNA editing result in multiple messenger RNA (mRNA) isoforms being generated from a single genomic locus. Most prevalently, alternative splicing is estimated to take place for over 90% of the multi-exon human genes [7], and thought to play critical roles in early stages of development and normal function of cells from diverse tissue types. Thus, the ability to reconstruct full length isoform sequences and accurately estimate their frequencies is critical for understanding gene functions and transcription regulation mechanisms.

High-throughput transcriptome sequencing, commonly referred to as RNA-Seq, is quickly replacing microarrays as the technology of choice for transcriptome analysis due to the far wider dynamic range and more accurate quantitation capabilities [8]. Unfortunately, most RNA-Seq studies to date either ignore alternative splicing or, similar to splicing array studies, restrict themselves to surveying the presence/expression levels of exons and exon-exon junctions. The main difficulty lies in the fact that current technologies used to perform RNA-Seq generate short reads (from few tens to hundreds of bases), many of which cannot be unambiguously assigned to individual isoforms.

In this abstract we introduce a novel EM algorithm for isoform frequency estimation from (any mixture of) single and paired RNA-Seq reads, under the assumption that a complete list of candidate isoforms is available. While current transcript libraries are still incomplete, we expect their coverage to increase rapidly. A key feature of our algorithm, referred to as IsoEM, is that it exploits a largely ignored source of disambiguation information provided by the distribution of insert sizes, which is typically tightly controlled during library preparation as recommended by sequencing instrument manufacturers. The recently published [6] is the only other work we are aware of that incorporates insert size distribution in conjunction with paired read data. We show that modeling insert sizes is also highly beneficial in conjunction with single RNA-Seq reads.

Insert sizes contribute to increased estimation accuracy in two different ways. On one hand, insert sizes are an important source of disambiguation information. In IsoEM, insert lengths are combined with base quality scores, and, if available, read pairing and strand information to probabilistically allocate reads

Dagstuhl Seminar Proceedings 10231
Structure Discovery in Biology: Motifs, Networks & Phylogenies
http://drops.dagstuhl.de/opus/volltexte/2010/2687

to isoforms during the expectation step of the EM algorithm. As in [2], the genomic locations of multireads are also resolved probabilistically in this step, further contributing to overall accuracy compared to methods that pre-select a unique genomic location by ad-hoc filtering rules. On the other hand, insert size distribution is used to accurately adjust isoform lengths during frequency re-estimation in the M step of the EM algorithm; an equivalent adjustment was independently employed in the probabilistic model of [6].

Preliminary experimental results on synthetic datasets generated with various sequencing parameters and distribution assumptions show that IsoEM algorithm significantly outperforms existing methods of isoform and gene expression level estimation from RNA-Seq data, including the widely used counting of unique reads, the multiread rescue method of [3], and the EM algorithms of [5] and [2]. Furthermore, we empirically evaluate the effect of sequencing parameters such as read length, read pairing, and strand information on estimation accuracy. We confirm the finding of [2] that, for a fixed total number of typed bases, longer reads are not necessarily better for estimation accuracy. In particular, for both single and paired read sequencing, 100bp reads are dominated by 50bp reads. This suggests that there may be limited benefits from further increases in read length and that higher sequencing depth is more critical to expression estimation accuracy.

Details on the IsoEM algorithm and experimental results are furthcoming in [4]. The open source Java implementation of IsoEM is available for download at http://dna.engr.uconn.edu/software/IsoEM/. In ongoing work we are integrating isoform frequency estimation with identification of novel transcripts using an iterative refinement framework similar to that proposed in [1].

Acknowledgments

This work was supported in part by NSF awards IIS-0546457, IIS-0916401, and IIS-0916948.

References

1. Wei Li Jianxing Feng and Tao Jiang. Inference of isoforms from short sequence reads. In *Proc. RECOMB*, 2010.
2. Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, February 2010.
3. Ali Mortazavi, Brian A A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 2008.
4. M. Nicolae, S. Mangul, I.I. Mandoiu, and A. Zelikovsky. Estimation of alternative splicing isoform frequencies from rna-seq data. In M. Singh and V. Moulton, editors, *Proc. 10th Workshop on Algorithms in Bioinformatics*, Lecture Notes in Computer Science, 2010 (to appear).

5. B. Paşaniuc, N. Zaitlen, and E. Halperin. Accurate estimation of expression levels of homologous genes in RNA-seq experiments. In *Proc. RECOMB*, pages 397–409, 2010.
6. Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, May 2010.
7. Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, November 2008.
8. Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, January 2009.